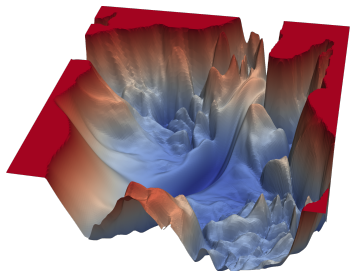


Introduction to Machine Learning

Advanced Risk Minimization Properties of Loss Functions



Learning goals

- Statistical properties
- Robustness
- Optimization properties
- Some fundamental terminology

THE ROLE OF LOSS FUNCTIONS

- Should be designed to measure errors appropriately
- **Statistical** properties: choice of loss implies statistical assumptions about the distribution of $y \mid \mathbf{x} = \tilde{\mathbf{x}}$
(see *maximum likelihood vs. empirical risk minimization*)
- **Robustness** properties:
some losses more robust towards outliers than others
- **Optimization** properties: computational complexity of

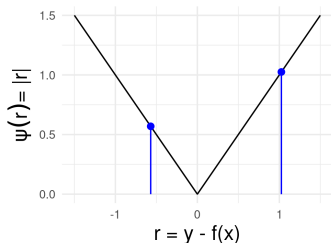
$$\arg \min_{\theta \in \Theta} \mathcal{R}_{\text{emp}}(\theta)$$

is influenced by choice of the loss

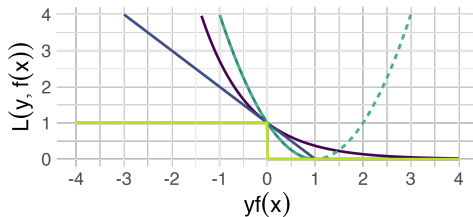


LOSSES WITH ONE ARGUMENT

- Regr. losses often only depend on **residuals** $r(\mathbf{x}) := y - f(\mathbf{x})$
- Classif. losses usually in terms of **margin**: $\nu(\mathbf{x}) := y \cdot f(\mathbf{x})$



Distance-based: L1

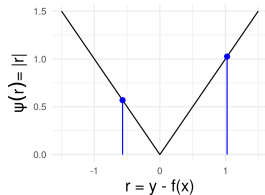


- Exponential
- Squared (scores)
- Hinge
- 0-1
- Squared hinge

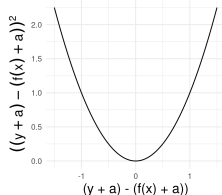
SOME BASIC PROPERTIES

A loss is

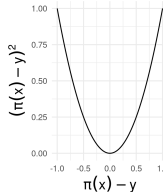
- **symmetric** if $L(y, f(\mathbf{x})) = L(f(\mathbf{x}), y)$
- **translation-invariant** if $L(y + a, f(\mathbf{x}) + a) = L(y, f(\mathbf{x}))$, $a \in \mathbb{R}$
- **distance-based** if it can be written in terms of residual
 $L(y, f(\mathbf{x})) = \psi(r)$ for some $\psi : \mathbb{R} \rightarrow \mathbb{R}$, and $\psi(r) = 0 \Leftrightarrow r = 0$



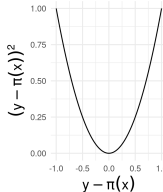
Distance-based: L1



Translation-invariant: L2



Symmetric: Brier score



ROBUSTNESS

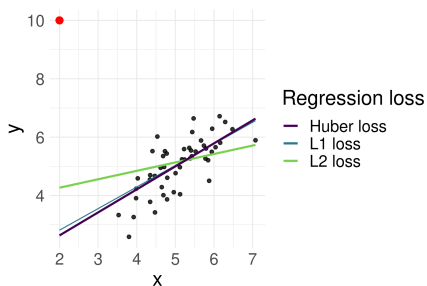
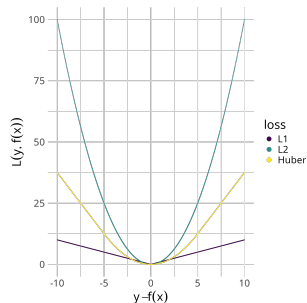
Outliers (in y) have large residuals $r(\mathbf{x}) = y - f(\mathbf{x})$. Some losses are more affected by large residuals than others. If loss goes up superlinearly (e.g. L2) it is not robust, linear (L1) or even sublinear losses are more robust.



$y - f(\mathbf{x})$	L1	L2	Huber ($\epsilon = 5$)
1	1	1	0.5
5	5	25	12.5
10	10	100	37.5
50	50	2500	237.5

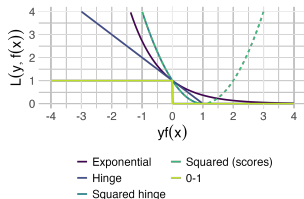
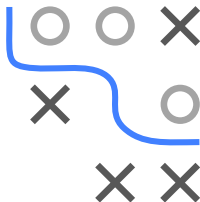
As a consequence, a model is less influenced by outliers than by “inliers” if the loss is **robust**.

Outliers e.g. strongly influence L2.



OPTIMIZATION PROPERTIES: SMOOTHNESS

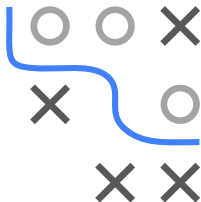
- Measured by number of continuous derivatives
- Usually want to have at least gradients in optimization of $\mathcal{R}_{\text{emp}}(\theta)$
- If loss is not differentiable, might have to use derivative-free optimization (or worse, in case of 0-1)
- Smoothness of $\mathcal{R}_{\text{emp}}(\theta)$ not only depends on L , but also requires smoothness of $f(\mathbf{x})$!



Squared loss, exponential loss and squared hinge loss are continuously differentiable.
Hinge loss is continuous but not differentiable.
0-1 loss is not even continuous.

OPTIMIZATION PROPERTIES: CONVEXITY

- Convexity of $\mathcal{R}_{\text{emp}}(\theta)$ depends both on convexity of $L(\cdot)$ (given in most cases) and $f(\mathbf{x} \mid \theta)$ (often problematic)
- If we model our data using an exponential family distribution, we always get convex losses ► Wedderburn 1976
- For $f(\mathbf{x} \mid \theta)$ linear in θ , linear/logistic/softmax/poisson/. . . regression are convex problems (all GLMs)!



Li et al., 2018: *Visualizing the Loss Landscape of Neural Nets*. The problem on the bottom right is convex, the others are not (note that very high-dimensional surfaces are coerced into 3D here).

